

AMiner: Search and Mining of Academic Social Networks

Huaiyu Wan¹, Yutao Zhang², Jing Zhang³ & Jie Tang^{2†}

¹ School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

² Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

³ Information School, Renmin University of China, Beijing 100872, China

Keywords: Academic social networks; Profile extraction; Name disambiguation; Topic modeling; Expertise Search; Network mining

Citation: H. Wan, Y. Zhang, J. Zhang & J. Tang. AMiner: Search and mining of academic social networks. Data Intelligence 1(2019), 58-76. doi: 10.1162/dint_a_00006

Received: May 22, 2018; Revised: June 8, 2018; Accepted: June 12, 2018

ABSTRACT

AMiner is a novel online academic search and mining system, and it aims to provide a systematic modeling approach to help researchers and scientists gain a deeper understanding of the large and heterogeneous networks formed by authors, papers, conferences, journals and organizations. The system is subsequently able to extract researchers' profiles automatically from the Web and integrates them with published papers by a way of a process that first performs name disambiguation. Then a generative probabilistic model is devised to simultaneously model the different entities while providing a topic-level expertise search. In addition, AMiner offers a set of researcher-centered functions, including social influence analysis, relationship mining, collaboration recommendation, similarity analysis and community evolution. The system has been in operation since 2006 and has been accessed from more than 8 million independent IP addresses residing in more than 200 countries and regions.

[†] Corresponding author: Jie Tang (Email: jietang@tsinghua.edu.cn; ORCID: 0000-0002-6882-4044).

1. INTRODUCTION

A variety of academic social networking websites including Google Scholar^①, Microsoft Academic^②, Semantic Scholar^③, ResearchGate^④ and Academia.edu^⑤ have gained great popularity over the past decade. The common purpose of these academic social networking systems is to provide researchers with an integrated platform to query academic information and resources, share their own achievements, and connect with other researchers.

Several issues within academic social networks have been investigated in these systems. However, most of the issues are investigated separately through independent processes. As such, there is not a congruent process or series of methods for mining the whole of disparate academic social networks. The lack of such methods can be attributed to two reasons:

- 1). Lack of semantic-based information. The user profile information obtained solely from the user who entered his or her information or extracted by heuristics is sometimes incomplete or inconsistent. Users do not fill in personal information merely because they are unwilling to do so;
- 2). Lack of a unified modeling approach for effective mining of the social network. Traditionally, different types of information sources in the academic social network were modeled individually, and thus dependencies between them cannot be captured. However, dependencies may exist between social data. High-quality search services need to consider the intrinsic dependencies between the different heterogeneous information sources.

AMiner[®] [1], the second generation of the ArnetMiner system, is designed to search and perform data mining operations against academic publications on the Internet, using social network analysis to identify connections between researchers, conferences and publications. In AMiner, our objective is to answer four questions:

- 1). How to automatically extract the researcher profile from the existing Web?
- 2). How to integrate the extracted information (i.e., researchers' profiles and publications) from different sources?
- 3). How to model the different types of information sources in a unified model?
- 4). How to provide powered search services in a constructed network?

^① <https://scholar.google.com/>

^② <https://academic.microsoft.com/>

^③ <https://www.semanticscholar.org/>

^④ <https://www.researchgate.net/>

^⑤ <https://www.academia.edu/>

^⑥ <https://www.aminer.cn/>

To answer the above questions, a series of novel approaches are implemented within the AMiner system. The overall architecture of the system is shown in Figure 1.

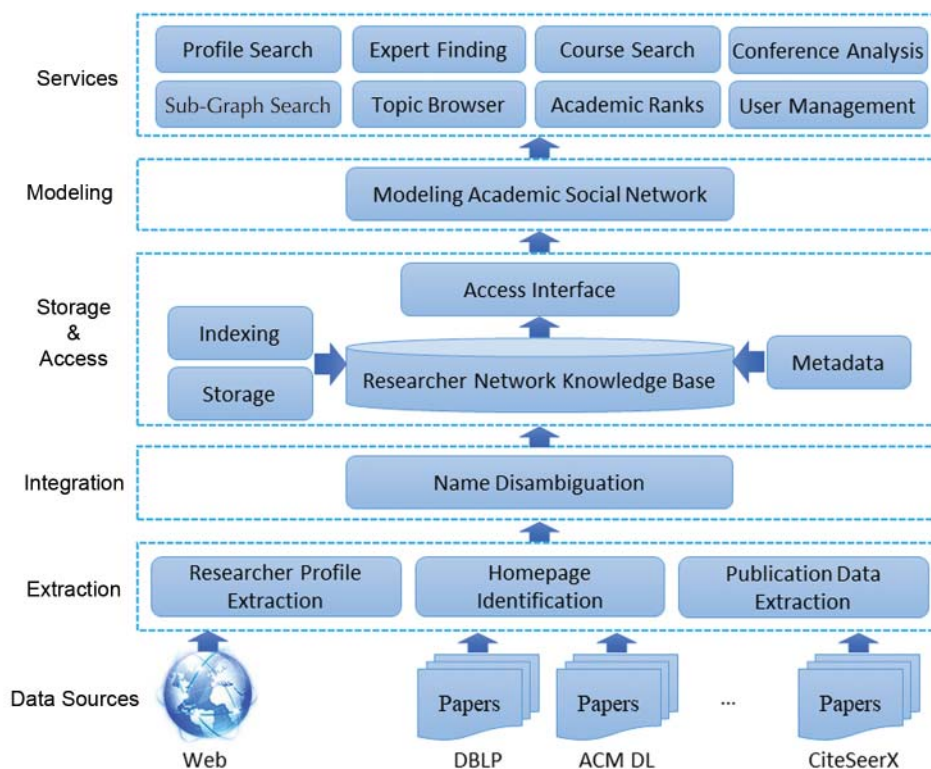


Figure 1. The architecture of AMiner.

The system mainly consists of five components:

1). Extraction. Focus is on automatically extracting researchers' profiles from the Web. The service first collects and identifies one's relevant pages (e.g., homepages or introducing pages) from the Web, then uses a unified approach [2, 3] to extract data from the identified documents. It also extracts publications from online digital libraries using heuristic rules. In addition, a simple but very effective approach is taken for profiling Web users by leveraging the power of big data [4].

2). Integration. Joining and integrating the extracted researchers' profiles and the extracted publications. The application employs the researcher name as the identifier. A probabilistic model [5] and a comprehensive framework [6] have been developed to deal with the name ambiguity problem in the integration. The integrated data are then stored, sorted and indexed into a researcher network knowledge base.

3). Storage and Access. Providing storage and indexing for the extracted and integrated data in the researcher network knowledge base. Specifically, for storage it employs Jena [7], a tool to store and retrieve ontological data; for indexing, it employs the inverted file indexing method, an established method for facilitating information retrieval [8].

4). Modeling. Utilizing a generative probabilistic model [1] to simultaneously model the different types of information sources. The system estimates a mixture of topic distribution associated with the different information sources.

5). Services. Providing several powered services based on the modeling results: profile search, expert finding, conference analysis, course search, sub-graph search, topic browser, academic ranks and user management.

For several features in the system, e.g., profile extraction, name disambiguation, academic topic modeling, expertise search and academic social network mining, we propose some new approaches to overcome the drawbacks that exist in the conventional methods.

The rest of this paper is organized as follows. Section 2 discusses related works, and Section 3 presents our proposed approaches in the system. Section 4 shows some applications of AMiner. Section 5 lists the data sets we constructed. Finally, Section 6 makes a conclusion.

2. RELATED WORK

Previously several issues in academic social network have been investigated and some systems were developed.

Google Scholar provides a search engine to identify the hyperlinks of publications that are publicly available or may be obtained through institutional libraries. Google Scholar is not a social networking website in the general sense, but it has become an important platform for searching academic resources, keeping up with the latest research, promoting one's own achievements, and tracking academic impact. Registered users can create a personal Google Scholar profile to post their research interests, manage their publications, correct their co-authors, and access their citations per year metrics. The social part of Google Scholar is very simple: a user can follow a researcher so that when he or she has a new publication or citation the user will receive an email; the user can also set up alerts based on his or her own research field.

Microsoft Academic [9] employs technologies of machine learning, semantic analysis and data mining to help users explore academic information more powerfully. A user can create an account and a public profile by claiming the publications he or she authored. Microsoft Academic provides more extensive "follow" functions. Users can follow researchers, publications, journals, conferences, organizations and research topics. Based on a user's publication history and the events the user is following, Microsoft Academic will show the most relevant items and news on his or her personalized homepage. In addition,

rather than providing a simple keyword-based search engine, Microsoft Academic presents relevant results and recommendations to help users discover more academic information resources of interest to support a more expansive learning and research experience.

Semantic Scholar is designed to be a “smart” search engine to help researchers find better academic publications faster. It uses a combination of machine learning, natural language processing and machine vision to analyze publications and extract important features, adding a supplementary layer of semantic analysis to the traditional methods of citation analysis. In comparison to Google Scholar and Microsoft Academic, Semantic Scholar can quickly highlight the most important papers and identify the connections between them. The resulting influential citations, images and key phrases that the engine provides quickly become more relevant and impactful to the user's work.

ResearchGate's aim and objective is to connect geographically distant researchers and allow them to communicate continuously. Registered users of the site each have a user profile and can share their research output including papers, data, book chapters, patents, research proposals, algorithms, presentations and software source code. Users can also follow the activities of others and engage in discussions with them. ResearchGate organizes itself mainly around research topics, and maintains its own index, i.e., the ResearchGate Score, based on the user's contribution to content, profile details and participation in interaction on the site. An example is asking questions and offering answers.

Academia.edu is a for-profit academic social networking website. It allows its users to create a profile, share their works, monitor their academic impact, select areas of interests and follow the research evolving in particular fields. Users can browse the networks of people with similar interests from around the world on the website. Academia.edu includes an analytics dashboard where users can see the influence and diffusion of their works in real time. In addition, Academia.edu has an alert service that sends registered users an email whenever a person whom they are following publishes a new paper. Academia.edu alerts anyone who is following a certain topic. In this way the awareness of a paper can be raised by potential citators through the alert system.

Although most of the above systems have integrated a gigantic amount of academic resources and provided abundant means of searching and querying social networking functions, they have not performed systematic semantic-level analysis or mining. Consequently, in our AMiner system, our primary objective is to provide a unified modeling approach to gaining a greater and deeper understanding of the semantic connection in large and heterogeneous academic networks consisting of authors, papers, conferences, journals and organizations. As a result, our system can provide topic-level expertise search and researcher-centered functions.

3. METHODOLOGY

In this section we introduce in detail the challenges we are addressing with our AMiner system, and we present our methods and solutions.

3.1 Profile Extraction

We define the schema of the researcher profile by extending the FOAF ontology [10], as shown in Figure 2. In the schema, 24 properties and two relations are defined [2, 3].

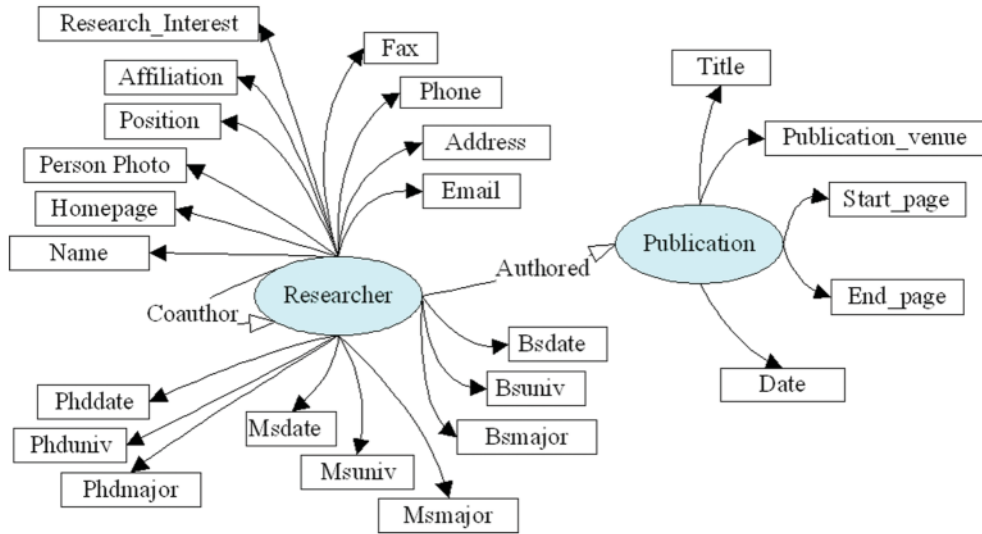


Figure 2. The schema of the researcher profile.

It is certainly not a trivial task to extract the research network from the Web. The researchers from different universities, institutes or companies have disparate page and profile templates and data feeds. So an ideal extraction method should consider processing all kinds of templates and formats. The approach we proposed consists of three steps:

1). Relevant page identification. Given a researcher name, we first get a list of Web pages by a search engine (Google API is used) and then identify the homepage or introducing page using a classifier. We define a set of features, such as whether the title of the page contains the person name and whether the URL address (partly) contains the person name, and employ Support Vector Machine (SVM) [11] for the classification.

2). Preprocessing. We separate the text into tokens and assign possible tags to each token. The tokens form the basic units and the pages form the sequences of the units in the following tagging step.

3). Tagging. Given a sequence of units, we determine the most likely corresponding sequence of tags by using a trained tagging model. The type of tags corresponds with the property defined in Figure 2. We define five types of tokens (i.e., standard word, special word, image token, term and punctuation mark) and use heuristics to identify tokens on the Web. After that, we assign several possible tags to each token based on the token type, and then a trained Conditional Random Fields (CRFs) model [12] is used to find the best tag assignment having the highest likelihood.

Recently, we revisit the problem of Web user profiling in the big data and propose a simple but very effective approach, referred to as MagicFG [4], for profiling Web users by leveraging the power of big data. To avoid error propagation, the approach integrates page identification and profile extraction in an unified framework. To improve the profiling performance, we present the concept of contextual credibility. The proposed framework also supports the incorporation of human knowledge. It defines human knowledge as Markov logics statements and formalizes them into a factor graph model. The MagicFG method has been deployed in AMiner system for profiling millions of researchers.

Figure 3 gives an example of researcher profile.

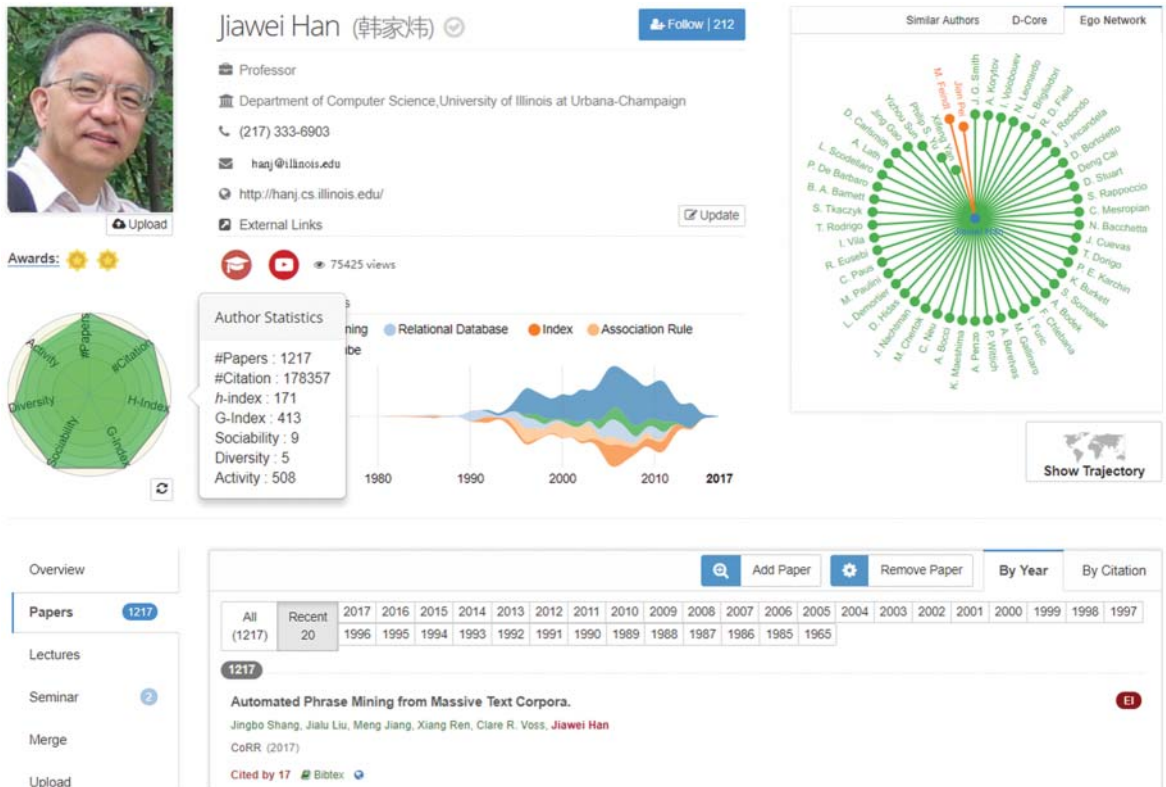


Figure 3. An example of researcher profile.

3.2 Name Disambiguation

We have collected more than 200 million publications from existing online data libraries, including DBLP[®], ACM DL[®], CiteSeerX[®] and others. In each data source, authors are identified by their names. For integrating the researcher profiles and the publication data, we use researcher name and the author name as the identifier. This process inevitably gives rise to an ambiguous result.

A few years ago, we proposed a probabilistic framework [5] based on Hidden Markov Random Fields (HMRF) [13] which is able to capture dependencies between observations (here each paper is viewed as an observation). The disambiguation problem is cast as assigning a tag to each paper with each tag representing an actual researcher.

More recently we proposed an additional comprehensive framework [6] to address the name disambiguation problem. The overview of the framework is shown in Figure 4. A novel representation learning method is proposed, which incorporates both global and local information. In addition, an end-to-end cluster size estimation method is presented in the framework. To improve the accuracy, we involve human annotators into the disambiguation process. The method has now been deployed in AMiner to deal with the name disambiguation problem at the billion scale, which demonstrates its effectiveness and efficiency.

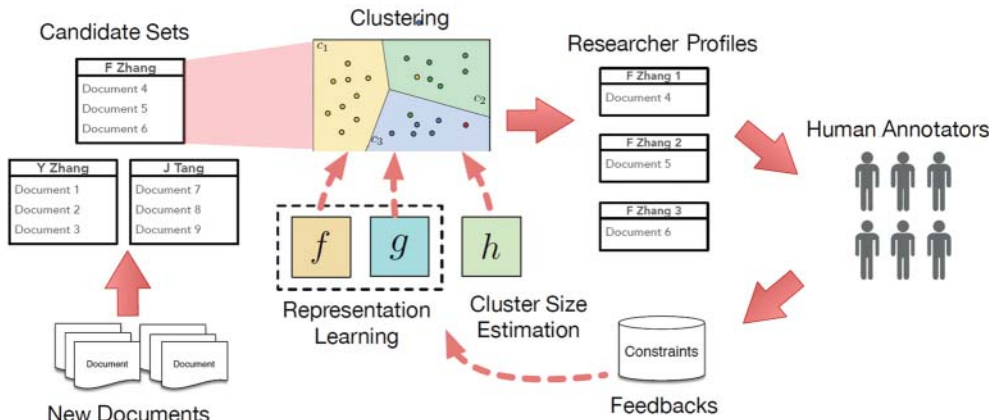


Figure 4. An overview of the name disambiguation framework in AMiner.

3.3 Topic Modeling

In academic search, representation of the content of text documents, author interests and conference themes are critical issues of any approach. Traditionally, documents are represented based on the “bag of words” (BOW) assumption. However, this representation cannot utilize the “semantic” dependencies between words. In addition, in the course of an academic search there are different types of information sources, and thus how to capture the dependencies between them becomes a challenging issue. Unfortunately, existing topic models, such as probabilistic Latent Semantic Indexing (pLSI) [14], Latent Dirichlet Allocation (LDA) [15] and Author-Topic model [16, 17] cannot be directly applied to the context

^② <https://dblp.uni-trier.de/>

^③ <https://dl.acm.org/>

^④ <http://citeseerx.ist.psu.edu/>

of academic search. This is because they simply cannot capture all intrinsic dependencies between papers and conferences.

A unified topic modeling approach [1] is proposed for simultaneously modeling characteristics of documents, authors, conferences and dependencies among them. (For simplicity, we use conference to denote conference, journal and book in the model.) The proposed model is called Author-Conference-Topic (ACT) model. More specifically, different strategies can be employed to model the topic distributions (as shown in Figure 5) and consequently the implemented models can have different knowledge representation capacities. In Figure 5(a) each author is associated with a mixture of weights over topics. For example, each word token correlated to a paper, and likewise a conference stamp associated to each word token, is generated from a sampled topic. In Figure 5(b) each author-conference pair is associated with a mixture of weights over the topics and word tokens are then generated from the sampled topics. In Figure 5(c), each author is associated with topics, each word token is generated from a sampled topic, and then a conference is generated from the sampled topics of all word tokens in a paper.

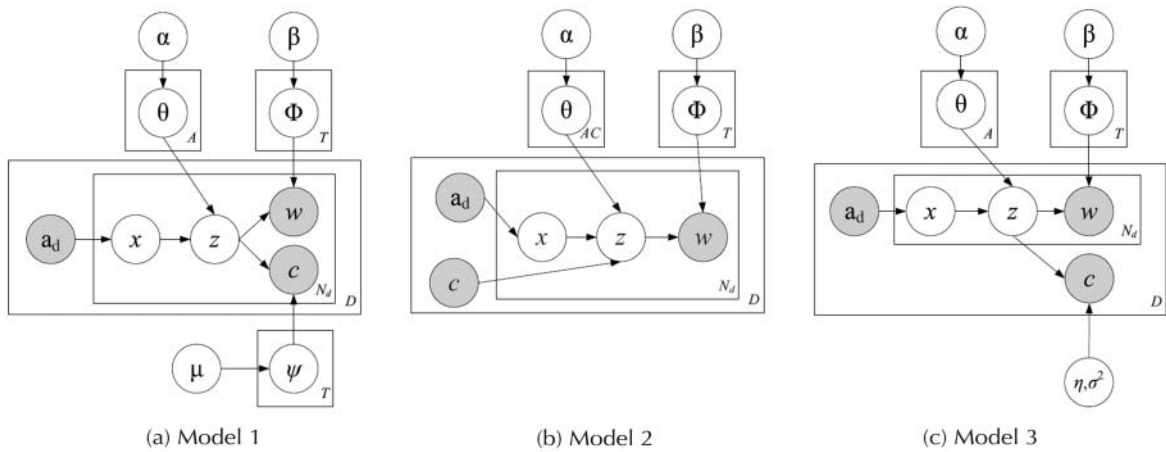


Figure 5. Graphical representation of the three Author-Conference-Topic (ACT) models.

3.4 Expertise Search

When searching for academic resources and formulating a query, a user endeavors to find authors with specific expertise, and papers and conferences related to the research areas of interest.

In the AMiner system we present a topic level expertise search framework [18]. Different from the traditional Web search engines that perform retrieval and ranking at document level, we study the expertise search problem at topic level over disparate heterogeneous networks. A unified topic model, namely Citation-Tracing-Topic (CTT), is proposed to simultaneously model topical aspects of different objects in the academic network. Based on the learned topic models, we investigate the expertise search problem from three

dimensions: ranking, citation tracing analysis and topic graph search. Specifically, we propose a topic level random walk method for ranking different objects. In citation tracing analysis we seek to uncover how a study influences its follow-up study. Finally, we have developed a topical graph search function, based on the topic modeling and citation tracing analysis.

Figure 6 gives an example result of experts found for the query “Data Mining”.

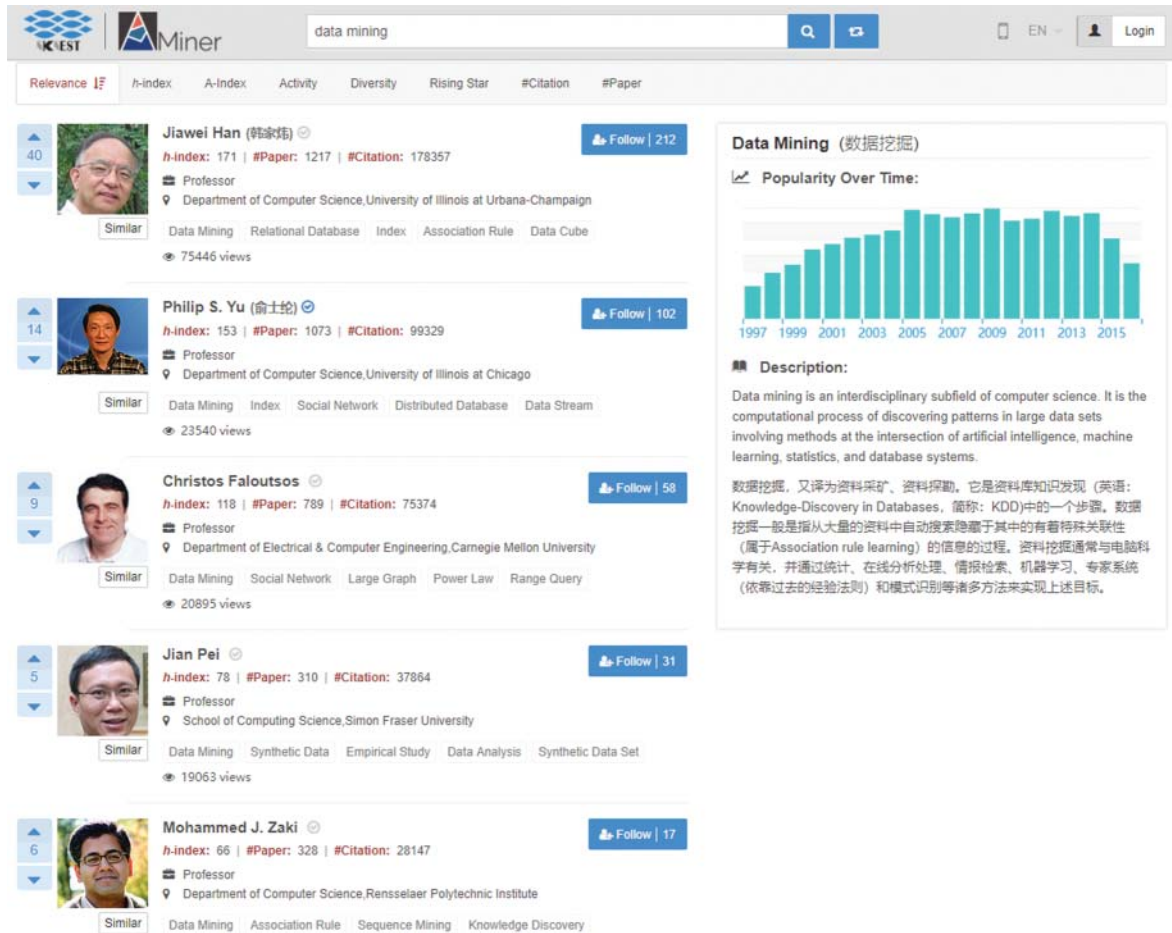


Figure 6. An example result of experts found for “Data Mining”.

3.5 Academic Social Network Mining

Based on the AMiner system, this set of researcher-centric academic social network mining functions includes social influence analysis, social relationship mining, similarity analysis, collaboration recommendation and community evolution.

Social Influence Analysis. In large social networks, persons are influenced by others for various reasons. We propose a Topic Affinity Propagation (TAP) model [19] to differentiate and quantify the social influence. TAP can take results of any topic modeling and the existing network structure to perform topic-level influence propagation. Recently, we design an end-to-end framework that we call DeepInf for feature representation learning and to predict social influence [20]. Each user is represented with a local sub-network which he or she is embedded in. A graph neural network is used to learn the representation of the sub-network which in turn effectively integrates the user-specific features and network structures. The framework of DeepInf is shown in Figure 7.

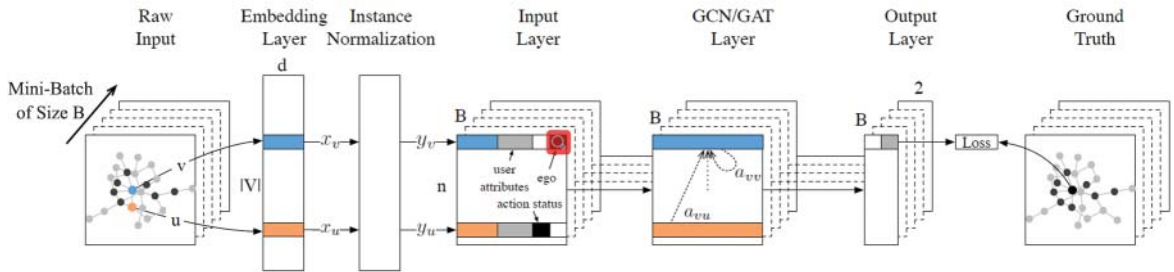


Figure 7. Model framework of DeepInf.

Social Relationship Mining. Inferring the type of social relationships between two users is a very important task in social relationship mining. We propose a two-stage framework named Time-constrained Probabilistic Factor Graph model (TPFG) [21] for inferring advisor-advisee relationships in the co-author network. The main idea is to leverage a time-constrained probabilistic factor graph model to decompose the joint probability of the unknown advisors over all the authors. Furthermore, we develop a framework named TranFG for classifying the type of social relationships across disparate heterogeneous networks [22]. The framework incorporates social theories into a factor graph model, which effectively improves the accuracy of predicting the types of social relationships in a target network by borrowing knowledge from another source network.

Similarity Analysis. Estimating similarity between vertices is a fundamental issue in social network analysis. We propose a sampling-based method to estimate the top- k similar vertices [23]. The method is based on the novel idea of the random path sampling method known as Panther. Given a particular network as a starting point, Panther randomly generates a number of paths of a pre-defined length, and then the similarity between two vertices can be modeled as estimating the possibility that the two vertices appear on the same paths.

Collaboration Recommendation. Interdisciplinary collaborations have generated a huge impact on society. However, it is usually hard for researchers to establish such cross-domain collaborations. We analyze the cross-domain collaboration data from research publications and propose a Cross-domain Topic Learning (CTL) model [24] for collaboration recommendation. For handling sparse connections, CTL consolidates the existing cross-domain collaborations through topic layers as opposed to utilizing

author layers. This alleviates the sparseness issue. For handling complementary expertise, CTL models topic distributions from source and target domains separately, as well as the correlation across domains. For handling topic skewness, CTL only models relevant topics to the cross-domain collaboration.

Community Evolution. Since social networks are rather dynamic, it is interesting to study how persons in the networks form different clusters and how the various clusters evolve over time. We study mining co-evolution of multi-typed objects in a special type of heterogeneous networks, called a star network. We subsequently examine how the multi-typed objects influence each other in the network evolution [25]. A Hierarchical Dirichlet Process Mixture Model-based evolution model is proposed which detects the co-evolution of a multi-typed objects in the form of multi-typed cluster evolution in dynamic star networks. An efficient inference algorithm is provided to learn the proposed model.

4. APPLICATION

AMiner is developed to provide comprehensive search and mining services for researcher social networks. In this system, we focus on: (1) creating a semantic-based profile for each researcher by extracting information from the distributed Web; (2) integrating academic data (e.g., the bibliographic data and the researcher profiles) from multiple sources; (3) accurately searching in the heterogeneous network; (4) analyzing and discovering interesting patterns from the built researcher social network. The main search and analysis functions in AMiner are summarized in the following section.

Profile Search. Input a researcher name (e.g., Jie Tang), the system will return the semantic-based profile created for the researcher using information extraction techniques. In the profile page, the extracted and integrated information include: contact information, photo, citation statistics, academic achievement evaluation, (temporal) research interest, educational history, personal social graph, research funding (currently only US and CN) and publication records (including citation information and the papers that are automatically assigned to several different domains).

Expert Finding. Input a query (e.g., data mining), the system will return experts on this topic. In addition, the system will suggest the top conference and the top ranked papers on this topic. There are two ranking algorithms: VSM and ACT. The former is similar to the conventional language model and the latter is based on our Author-Conference-Topic (ACT) model. Users can also provide feedbacks to the search results.

Conference Analysis. Input a conference name (e.g., KDD), the system returns those who are the most active researchers on this conference as well as the top-ranked papers.






Course Search. Input a query (e.g., data mining), the system will return those who are teaching courses relevant to the query.

Sub-Graph Search. Input a query (e.g., data mining), the system first tells you what topics are relevant to the query (e.g., five topics “Data Mining”, “XML Data”, “Data Mining/Query Processing”, “Web Data/Database Design” and “Web Mining” are relevant), and subsequently display the most important sub-graph discovered on each relevant topic, augmented with a summary for the sub-graph.

Topic Browser. Based on our Author-Conference-Topic (ACT) model, we automatically discover 200 hot topics from the publications. For each topic, we automatically assign a label to represent its meanings. Furthermore, the browser presents the most active researchers, the most relevant conferences/papers, and the evolution trend of the topics that are discovered.

Academic Ranks. We define eight measures to evaluate the researcher's achievement. The measures include "h-index", "Citation", "Uptrend", "Activity", "Longevity", "Diversity", "Sociability" and "New Star". For each measure, we output a ranking list in different domains. For example, one can search those who have the highest citation numbers in the "data mining" domain. Figure 8 gives an example of researcher ranking by sociability index.

Researcher Rank

Overall h-Index G-Index #Citation #Paper Sociability Diversity						
					<u>Sociability</u>	<u>Rank</u>
	J. Huston h-index: 124 #Paper: 2193 #Citation: 121371 Professor Department of Physics and Astronomy/Michigan State University Search For Cross Sections Cross Section Standard Model Large Hadron Collider				10.943	1
	G. Mitselmakher h-index: 163 #Paper: 1963 #Citation: 198165 Professor Department of Physics University of Florida Search For Gravitational Waves Black Holes Cross Sections Standard Model				10.89	2
	A. Soffer h-index: 93 #Paper: 2387 #Citation: 85815 Professor Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel Search For T And And B Cross Sections Upper Limit				10.855	3
	A. Warburton h-index: 113 #Paper: 2127 #Citation: 96192 Associate Professor Department of Physics McGill University Search For Cross Sections Cross Section Standard Model And B				10.825	4
	W. Walkowiak h-index: 86 #Paper: 1087 #Citation: 85517 scientific staff member Fachbereich Physik Universität Siegen Cosmic Ray Cosmic Rays Charged Particles Charged Particle				10.82	5

HELP

Experts' Statistics

We calculate several features of authors, including h-index, A-Index, G-Index, Total citation number, Diversity, Sociability, Activity, New Star and Rising Star. [please click here.](#)

If you find a bug, please [sent email to us.](#)

Figure 8. An example of researcher ranking by sociability index.

User Management. One can register as a user to: (1) modify the extracted profile information; (2) provide feedback on the search results; (3) follow researchers in AMiner; and (4) to create an AMiner page (which can be used to advertise conferences and workshops, or to recruit students).

5. DATA SET

AMiner has collected a large scholar data set with more than 130,000,000 researcher profiles and 233,000,000 publications from the Internet by June 2018 along with a number of subsets that were constructed for different research purposes. The details of these subsets are as follows, and can be found at <https://www.aminer.cn/data>.

Citation Network. The citation data are extracted from DBLP, ACM DL and other sources. The data set contains 1,572,277 papers and 2,084,019 citation relationships. Each paper is associated with abstract, authors, year, venue and title. The data set can be used for clustering with network and side information, studying influence in the citation network, finding the most influential papers, topic modeling analysis, etc.

Academic Social Network. These data include papers, paper citation, author information and author collaboration. The data set contains 1,712,433 authors, 2,092,356 papers, 8,024,869 citation relationships and 4,258,615 collaboration relationships noted between authors.

Advisor-Advisee. The data set is comprised of 815,946 authors and 2,792,833 co-author relationships. For evaluating the performance of inferring advisor-advisee relationships between co-authors, we created a smaller ground truth data using the following method: (1) collecting the advisor-advisee information from the Mathematics Genealogy project and the AI Genealogy project; (2) manually crawling the advisor-advisee information from researchers' homepages. Finally, we have labeled 1,534 co-author relationships, of which 514 are advisor-advisee relationships.

Topic-Co-Author. It is a topic-based co-author network, which contains 640,134 authors of eight topics and 1,554,643 co-author relationships. The eight topics are: Data Mining/Association Rules, Web Services, Bayesian Networks/Belief Function, Web Mining/Information Fusion, Semantic Web/Description Logics, Machine Learning, Database Systems/XML Data and Information Retrieval.

Topic-Paper-Author. The data set is collected for the purpose of cross domain recommendation which contains 33,739 authors associated to five topics as well as 139,278 co-author relationships. The five topics are Data Mining (with 6,282 authors and 22,862 co-author relationships), Medical Informatics (with 9,150 authors and 31,851 co-author relationships), Theory (with 5,449 authors and 27,712 co-author relationships), Visualization (with 5,268 authors and 19,261 co-author relationships) and Database (with 7,590 authors and 37,592 co-author relationships).

Topic-Citation. It is a topic-based citation network, which contains 2,329,760 papers of 10 topics and 12,710,347 citations relationships. The 10 topics are: Data Mining/Association Rules, Web Services, Bayesian Networks/Belief Function, Web Mining/Information Fusion, Semantic Web/Description Logics, Machine Learning, Database Systems/XML Data, Pattern Recognition/Image Analysis, Information Retrieval and Natural Language System/Statistical Machine Translation.

Kernel Community. It is a co-authorship network with 822,415 nodes and 2,928,360 undirected edges. Each vertex represents an author and each edge represents a co-author relationship.

Dynamic Co-Author. The data set contains 1,768,776 papers published during the time period from 1986 to 2012 with 1,629,217 authors involved. Each year is regarded as a time stamp and there are 27 time stamps in total. At each time stamp, we create an edge between two authors if they have co-authored at least one paper in the most recent three years (including the current year). We convert the undirected co-author network into a directed network by regarding each undirected edge as two symmetric directed edges.

Expert Finding. This data set is a benchmark for expert finding, which contains 1,781 experts of 13 topics.

Association Search. This data set is used to evaluate the effectiveness of association search approaches, which contains 8,369 author pairs specific to nine topics. Each author pair contains a source author and target author.

Topic Model Results for AMiner Data Set. There are the results of ACT model on the AMiner data set which contains the top 1,000,000 papers and authors of 200 topics.

Co-Author. This is a co-author network on the AMiner system which contains 1,560,640 authors and 4,258,946 co-author relationships.

Disambiguation. This data set is used for studying name disambiguation in a digital library. It contains 110 authors and their affiliations as well as their disambiguation results (ground truth).

6. CONCLUSION

In this paper, we present a novel online academic searching and mining system, AMiner. It is the second generation of the ArnetMiner system. We first present the overall architecture of the system, which consists of five main components, i.e., extraction, integration, storage and access, modeling and services. Then we follow this by introducing the important methodologies proposed in the system, including the profile extraction and user profiling methods, name disambiguation algorithms, topic modeling methods, expertise search strategies and series of academic social network mining methods. Furthermore, we introduce the typical applications as well as a broad and significant offering of available data sets already presented on the platform.

We acknowledge that AMiner is still at its developmental stage on both the scale of resources and the quality of services. However, in the future we are going to exploit additional intelligent methods for mining deep knowledge from scientific networks and we will deploy a more convenient and personalized framework for delivering academic search and finding services.

AUTHOR CONTRIBUTIONS

This work was a collaboration between all of the authors. J. Tang (jietang@tsinghua.edu.cn) is the leader of the AMiner project, who drew the whole picture of the system. Y. Zhang (yt-zhang13@mails.tsinghua.edu.cn) and J. Zhang (zhang-jing@ruc.edu.cn) summarized the methodology part of this paper. H. Wan (hywan@bjtu.edu.cn) summarized the applications and data sets in the AMiner system and drafted the paper. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

REFERENCES

- [1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, & Z. Su. ArnetMiner: Extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08), 2008, pp. 990–998. doi: 10.1145/1401890.1402008.
- [2] J. Tang, D. Zhang, & L. Yao. Social network extraction of academic researchers. In: Proceedings of 2007 IEEE International Conference on Data Mining (ICDM'07), 2007, pp. 292–301. doi: 10.1109/ICDM.2007.30.
- [3] J. Tang, L. Yao, D. Zhang, & J. Zhang. A combination approach to Web user profiling. ACM Transactions on Knowledge Discovery from Data 5(1) 2010, Article No. 2. doi: 10.1145/1870096.1870098.
- [4] X. Gu, H. Yang, J. Tang, J. Zhang, F. Zhang, D. Liu, W. Hall, & X. Fu. Profiling Web users using big data. Social Network Analysis and Mining 8(1) 2018, Article No. 24. doi: 10.1007/s13278-018-0495-0.
- [5] J. Tang, A.C.M. Fong, B. Wang, & J. Zhang. A unified probabilistic framework for name disambiguation in digital library. IEEE Transaction on Knowledge and Data Engineering 24(6) 2012, 975–987. doi: 10.1109/TKDE.2011.13.
- [6] Y. Zhang, F. Zhang, P. Yao, & J. Tang. Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18), 2018, pp. 1002–1011. doi: 10.1145/3219819.3219859.
- [7] J.J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, & K. Wilkinson. Jena: Implementing the semantic Web recommendations. In: Proceedings of the 13th World Wide Web Conference (WWW'04), 2004, pp. 74–83. doi: 10.1145/1013367.1013381.
- [8] C.J. van Rijsbergen. Information retrieval. London: Butterworths, 1979.
- [9] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.J. Hsu, & K. Wang. An overview of Microsoft Academic Service (MA) and applications. In: Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), 2015, pp. 243–246. doi: 10.1145/2740908.2742839.
- [10] D. Brickley, & L. Miller. FOAF vocabulary specification. Available at: <http://xmlns.com/foaf/0.1/>.
- [11] C. Cortes, & V. Vapnik. Support-vector networks. Machine Learning 20(3)(1995), 273–297. doi: 10.1007/BF00994018.
- [12] J.D. Lafferty, A. McCallum, & F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning (ICML'01), 2001, pp. 282–289. Available at: <http://portal.acm.org/citation.cfm?id=655813>.
- [13] S. Basu, M. Bilenko, & R.J. Mooney. A probabilistic framework for semi-supervised clustering. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004, pp. 59–68. doi: 10.1145/1014052.1014062.
- [14] T. Hofmann. Probabilistic Latent Semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999, pp. 50–57. doi: 10.1145/312624.312649.

- [15] D.M. Blei, & J.D. McAuliffe. Supervised topic models. In: Proceedings of the 19th Neural Information Processing Systems (NIPS'07), 2007, pp. 121–128. Available at: <http://papers.nips.cc/paper/3328-supervised-topic-models>.
- [16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, & P. Smyth. The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04), 2004, pp. 487–494. Available at: <https://dl.acm.org/citation.cfm?id=1036902>.
- [17] M. Steyvers, P. Smyth, & T. Griffiths. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004, pp. 306–315. doi: 10.1145/1014052.1014087.
- [18] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, & Z. Su. Topic level expertise search over heterogeneous networks. Machine Learning Journal 82(2)(2011), 211–237. doi: 10.1007/s10994-010-5212-9.
- [19] J. Tang, J. Sun, C. Wang, & Z. Yang. Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09), 2009, pp. 807–816. doi: 10.1145/1557019.1557108.
- [20] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, & J. Tang. DeepInf: Modeling influence locality in large social networks. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18), 2018, pp. 2110–2119. Available at: <https://www.haoma.io/pdf/deepinf.pdf>.
- [21] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, & J. Guo. Mining advisor-advisee relationships from research publication networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), 2010, pp. 203–212. doi: 10.1145/2910896.2925435.
- [22] J. Tang, T. Lou, J. Kleinberg, & S. Wu. Transfer learning to infer social ties across heterogeneous networks. ACM Transactions on Information Systems 34(2)(2016), Article No. 7. doi: 10.1145/2746230.
- [23] J. Zhang, J. Tang, C. Ma, H. Tong, Y. Jing, & J. Li. Panther: Fast top-k similarity search on large networks. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), 2015, pp. 1445–1454. doi: 10.1145/2783258.2783267.
- [24] J. Tang, S. Wu, J. Sun, & H. Su. Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), 2012, pp. 1285–1293. Available at: <http://keg.cs.tsinghua.edu.cn/jietang/publications/KDD12-Tang-et-al-Cross-Domain-Collaboration-Recommendation.pdf>.
- [25] Y. Sun, J. Tang, J. Han, C. Chen, & M. Gupta. Co-evolution of multi-typed objects in dynamic star networks. IEEE Transaction on Knowledge and Data Engineering 26(12)(2014), 2942–2955. doi: 10.1109/TKDE.2013.103.

AUTHOR BIOGRAPHY

Huaiyu Wan is an Associate Professor at the Department of Computer Science, Beijing Jiaotong University. He received his PhD degree from School of Computer and Information Technology, Beijing Jiaotong University. His current research interests include social network mining, user behavior analysis and traffic data mining.



Yutao Zhang is a postdoctoral researcher at the Department of Computer Science and Technology, Tsinghua University. He received his PhD degree from the Department of Computer Science and Technology, Tsinghua University. His current research interests include social network mining, text mining and visual analytics.



Jing Zhang is an Assistant Professor at the Department of Computer Science and Technology, Information School, Renmin University of China. She received her PhD degree from the Department of Computer Science and Technology, Tsinghua University. Her current research interests include social network mining, graph mining, text mining and deep learning.



Jie Tang is an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. His main research interests include data mining algorithms and social network theories. He has been a visiting scholar with Cornell University, Chinese University of Hong Kong, Hong Kong University of Science and Technology and Leuven University. He has published more than 100 research papers in major international journals, such as *Machine Learning*, *ACM Transactions on Knowledge Discovery from Data* (TKDD) and *IEEE Transactions on Knowledge and Data Engineering* (TKDE) and conferences including: the Knowledge Discovery and Data Mining (KDD) conference, the International Joint Conference on Artificial Intelligence (IJCAI), the AAAI Conference on Artificial Intelligence, the International Conference on Machine Learning, the International World Wide Web Conference (WWW), the ACM SIGIR Conference on Research and Development in Information Retrieval, the Conference of the ACM Special Interest Group on Data Communication (SIGMOD), and the Annual Meeting of the Association for Computational Linguistics (ACL).